Scientific Data Management Center

(Integrated Software Infrastructure Center – ISIC)

Arie Shoshani, LBNL

SciDAC PI Meeting
March, 2003

(http://sdmcenter.lbl.gov)

Participating Institutions

Center Director: Arie Shoshani

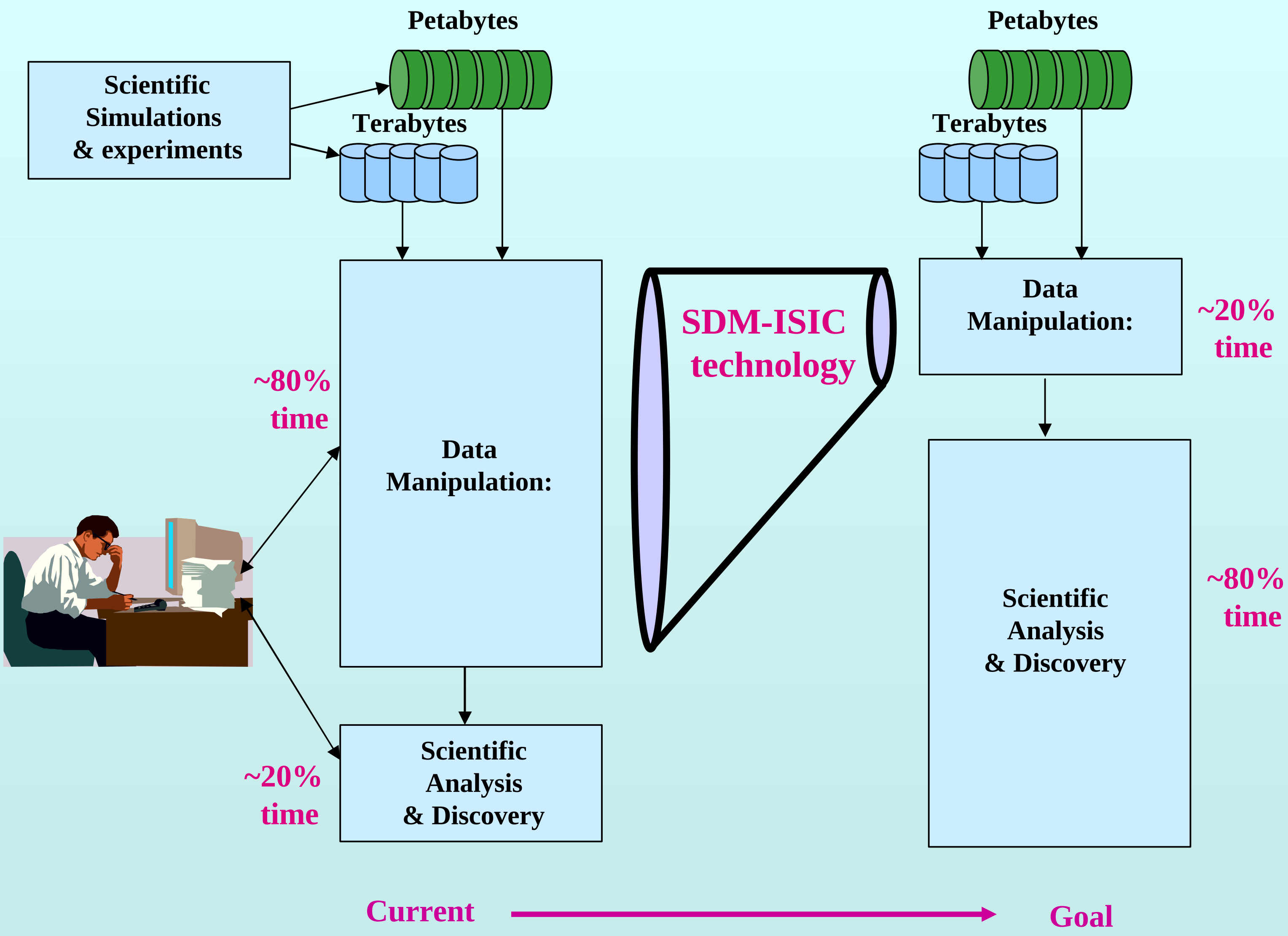
DOE Laboratories co-ordinating PIs:

ANL:	**Bill Gropp**
LBNL:	**Arie Shoshani**
LLNL:	**Terence Critchlow**
ORNL:	**Thomas Potok**

Universities co-ordinating PIs :

Georgia Institute of Technology: Calton Pu
North Carolina State University: Mladen Vouk
Northwestern University: Alok Choudhary
UC San Diego (Supercomputer Center): Reagan Moore

SDM Center Goal:
Reduce the Data Management Overhead



Scientific Simulations & experiments

Petabytes

Terabytes

~80% time

Data Manipulation:

~20% time

Scientific Analysis & Discovery

SDM-ISIC technology

Petabytes

Terabytes

Data Manipulation:

~20% time

Scientific Analysis & Discovery

~80% time

Current

Goal

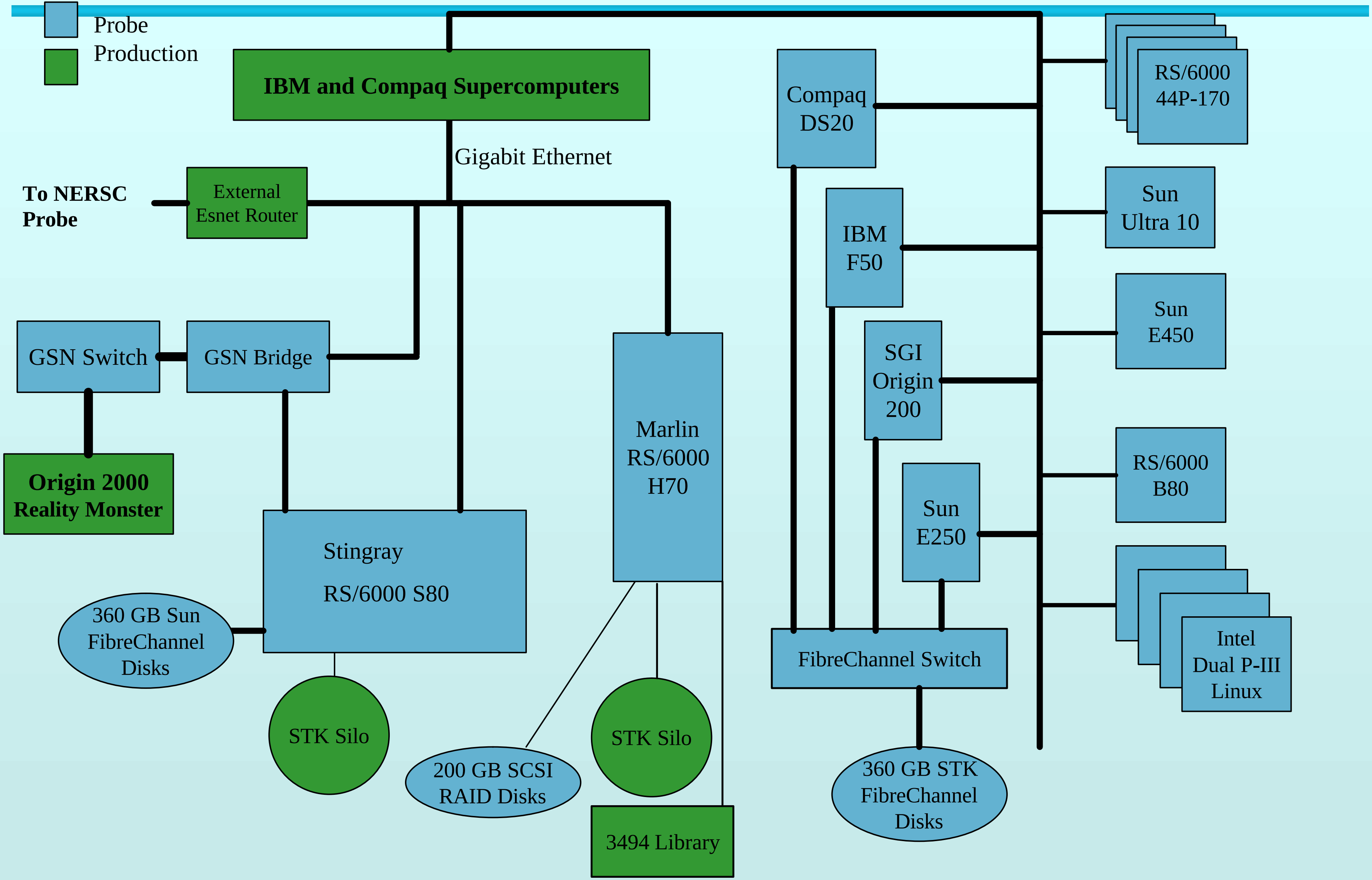
Reduce the Data Management Overhead: How?

- Efficiency
  - Example: parallel I/O, indexing, matching storage structures to the application
- Effectiveness
  - Example: Access data by attributes-not files, facilitate massive data movement
- New algorithms
  - Example: Specialized PCA techniques to separate signals or to achieve better spatial data compression
- Enabling ad-hoc exploration of data
  - Example: by enabling exploratory “run and render” capability to analyze and visualize simulation output while the code is running

Principles

- Guiding principles
  - Work with individual application scientists
  - Work with specific scientific problems
  - Deploy technology already developed or prototyped
  - Do research/development driven by need and experience
  - Re-apply techniques to new applications
- Focus areas
  - Parallel and Grid I/O Infrastructure
    - Astrophysics, Climate, Fusion
  - Exploratory Analysis and Data Mining
    - Astrophysics, Climate
  - Distributed, Heterogeneous Data Integration
    - Biology
  - Efficient Processing and Access of Very Large Datasets
    - High Energy Physics, Combustion, Astrophysics

Test Environment: ORNL Probe – “Place to be”
Randy Burris, ORNL

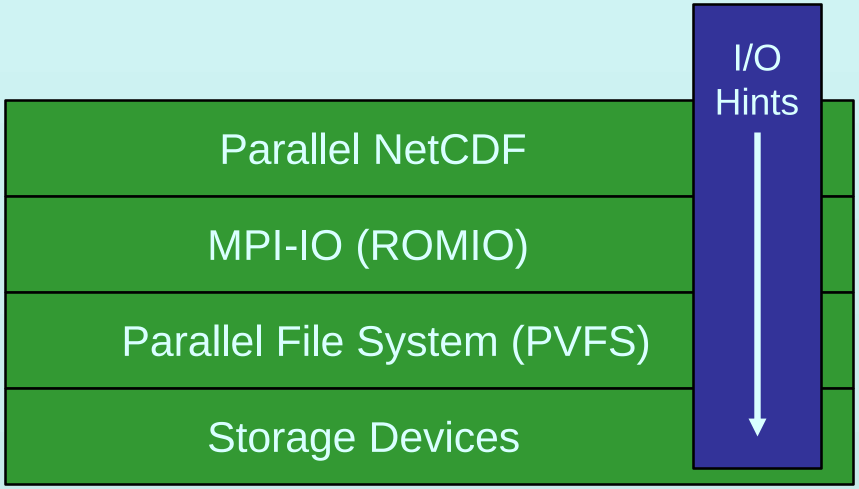


Results

(Also see 5 posters)

Parallel and Grid I/O

- Team Members
  - Bill Gropp, Rob Ross, Rajeev Thakur, Rob Latham, Neill Miller (ANL)
  - Alok Choudhary, Wei-Keng Liao, Jianwei Li, Avery Ching (NWU)
  - Ghaleb Abdulla, Tina Eliassi-Rad (LLNL)
- Improving software at all I/O layers
  - High level interfaces
  - MPI-IO
  - Parallel file systems
- Enabling tighter coupling of layers
  - Hints
  - Structured I/O requests

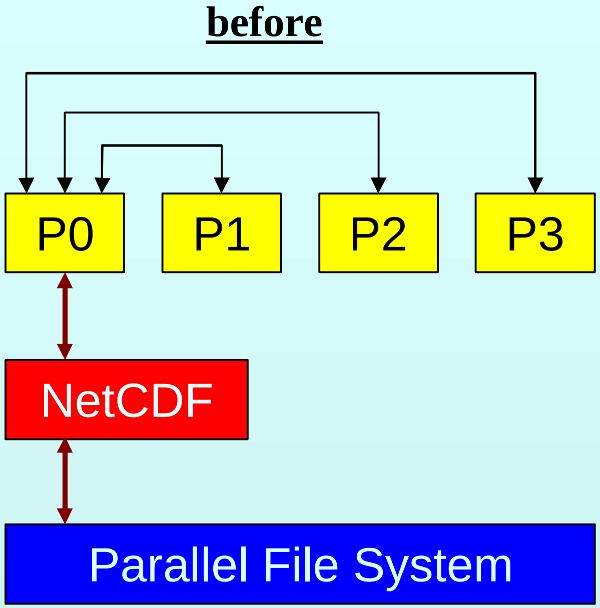


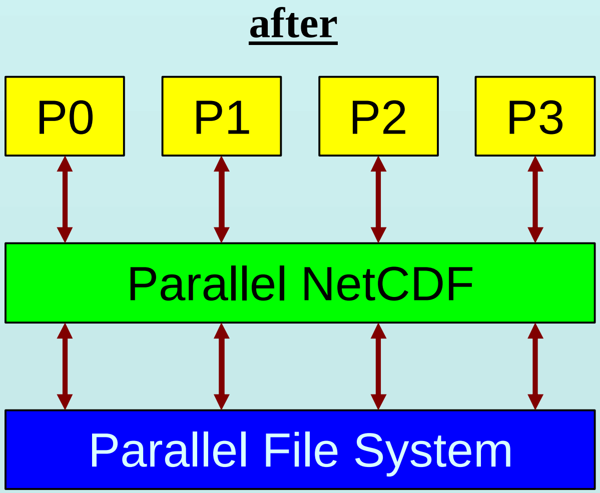
I/O in NetCDF

- **Original NetCDF**
  - No possibility of collective optimizations
  - All processes read file independently
  - Writes are carried out by shipping data to a single process (sequential write)

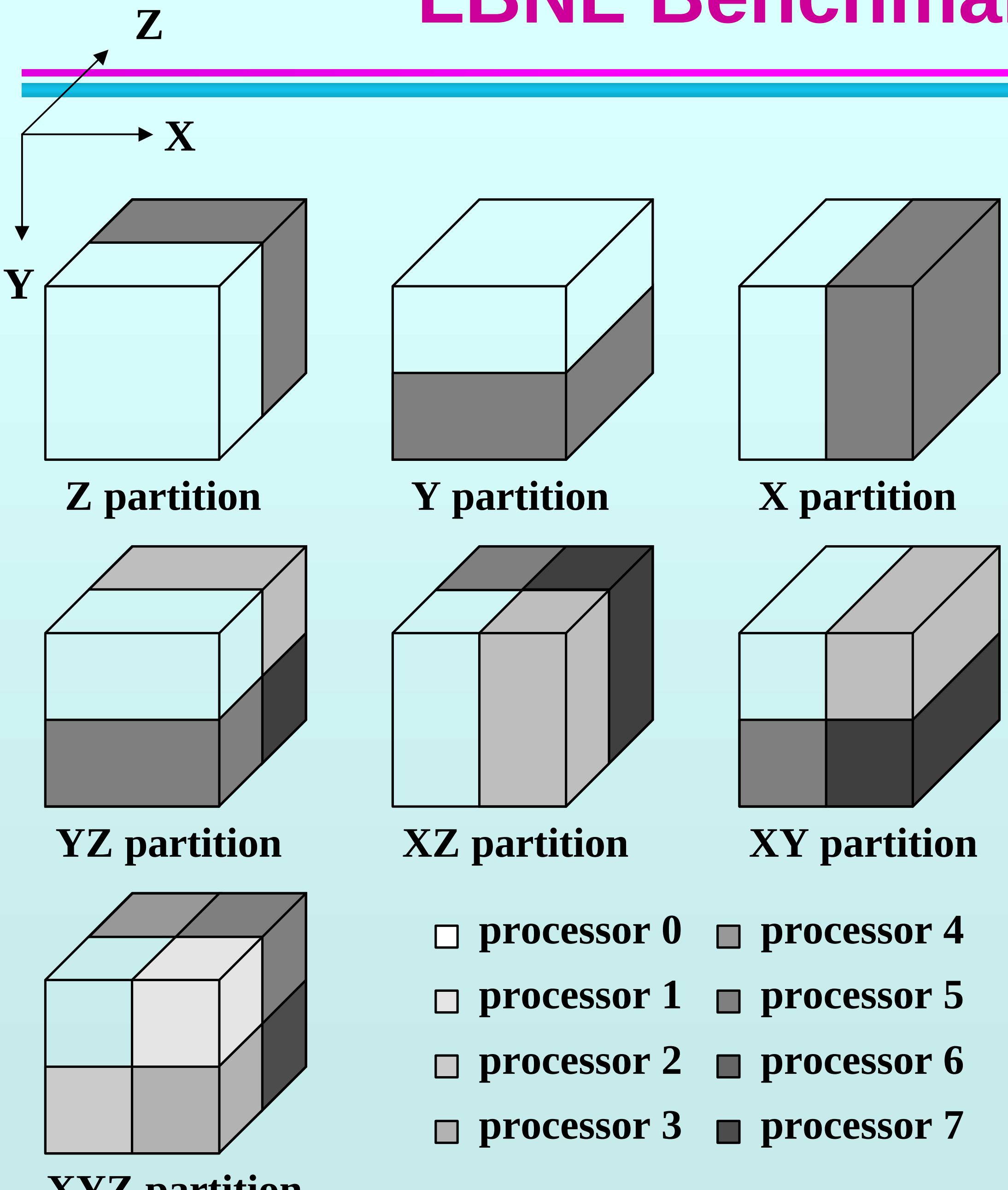
Parallel NetCDF

- Parallel read/write to shared NetCDF file
- Built on top of MPI-IO which utilizes optimal I/O facilities provided by the parallel file systems
- Allows for MPI-IO collective I/O, datatypes, and hints for further optimization



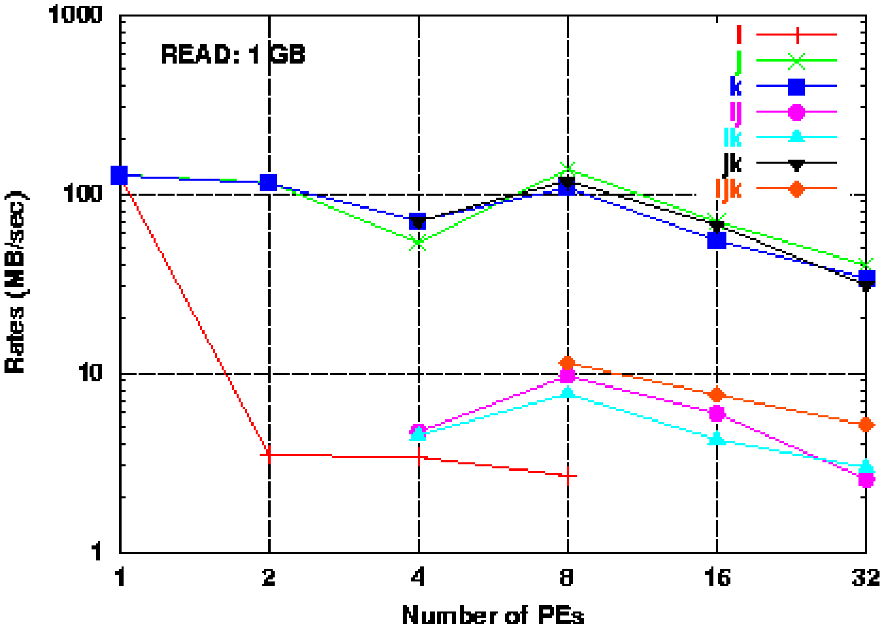


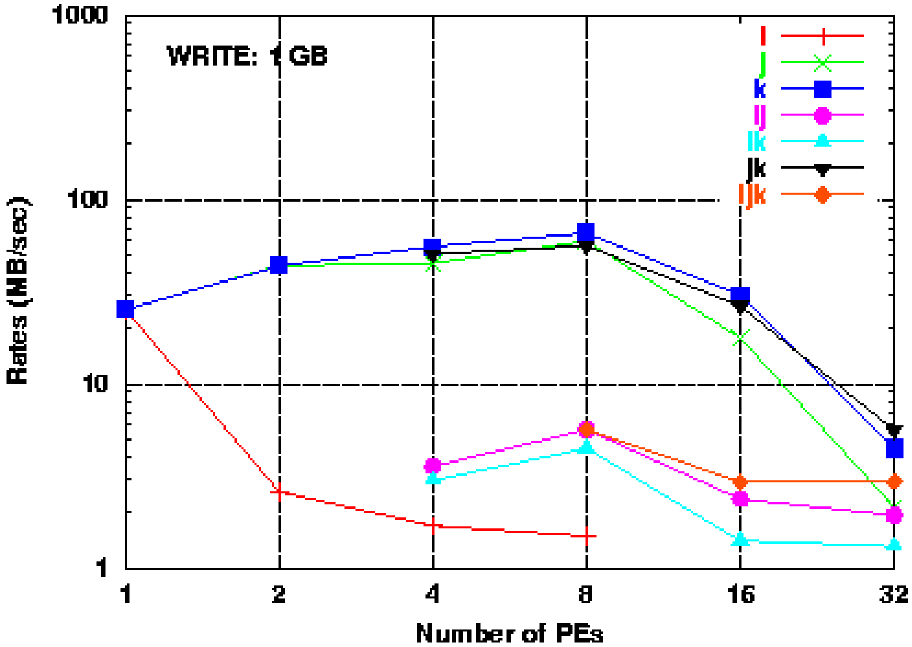
LBNL Benchmark



- **Test suite**
  - **Developed by Chris Ding et al. at LBNL**
  - **Written in Fortran**
  - **Simple block partition patterns**
- **Access to a 3D array which is stored in a single netCDF file**
- **Running on IBM SP2 at NERSC, LBNL**
  - **Each compute node is an SMP with 16 processors**
  - **I/O is performed using all processors**

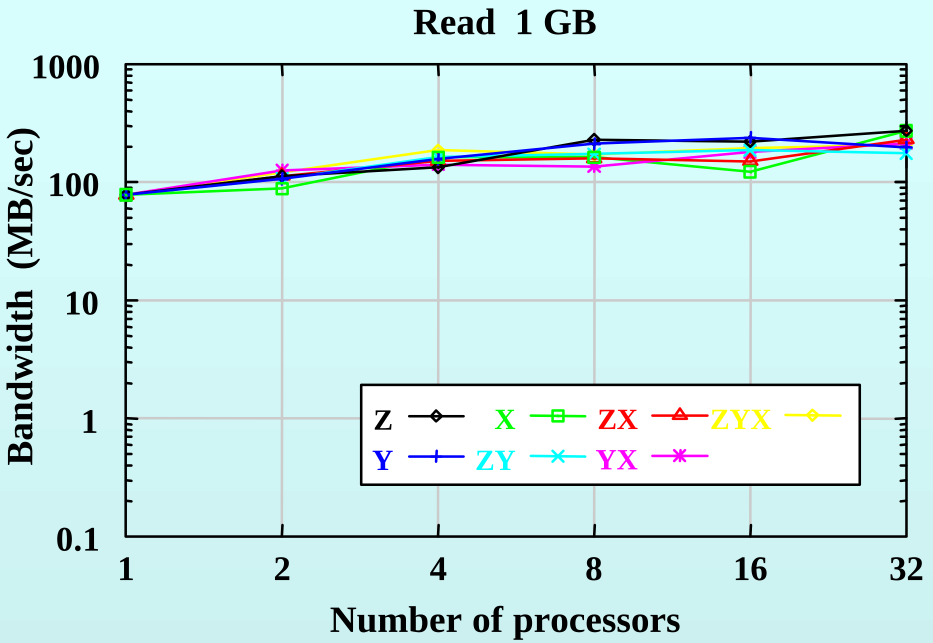
LBNL Results – 1 GB

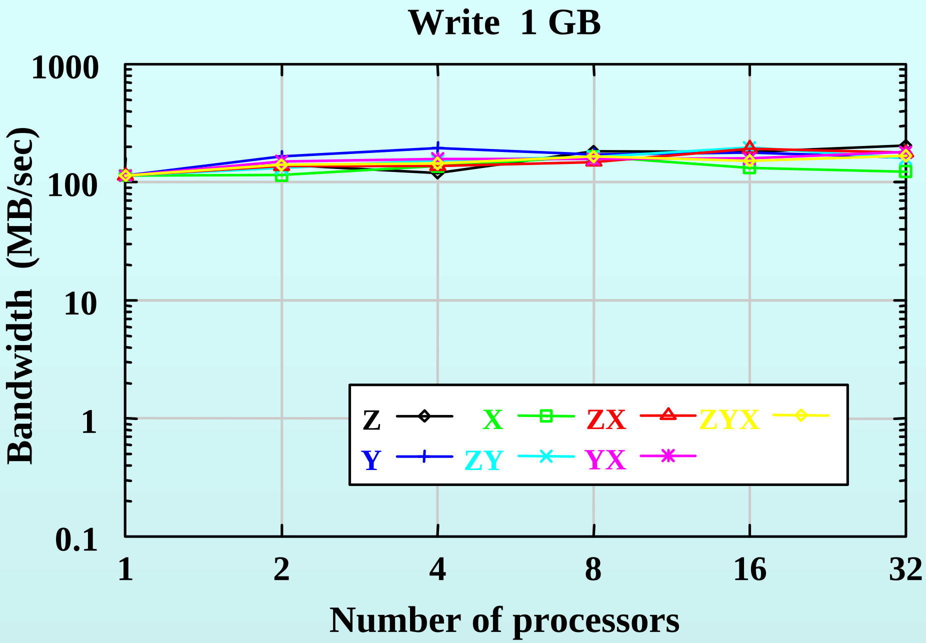




- Array size – 512 x 512 x 512, real*8
- Read
  - No better performance is observed
- Write
  - 4-8 processor writes results in 2-3 times higher bandwidth than using a single processor

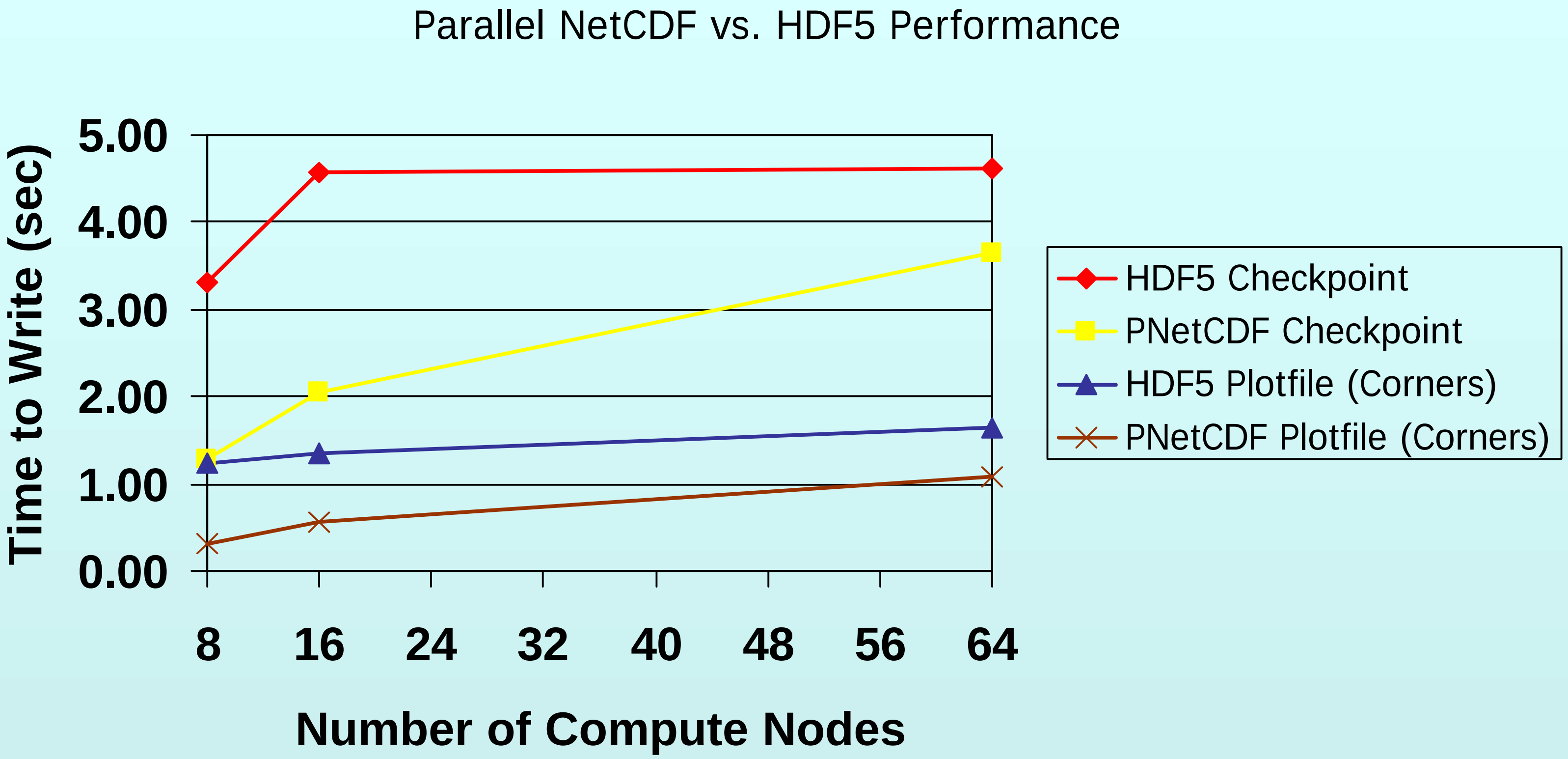
Our Results – 1 GB





- Array size: 512 x 512 x 512, real*8
- Run on IBM SP2 at SDSC
- I/O is performed using one processor per node

FLASH I/O Benchmark Comparison



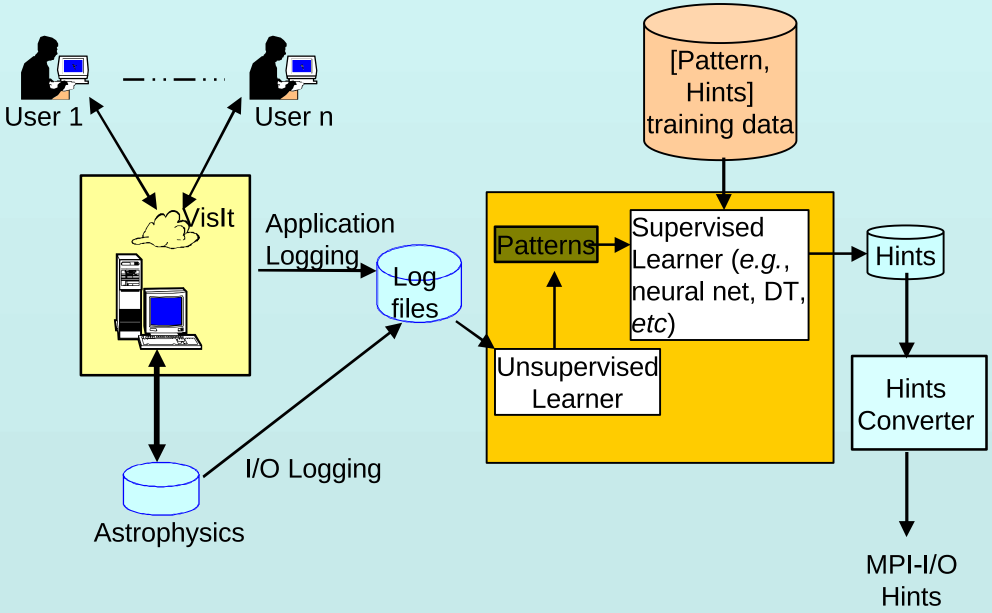
- Ported FLASH I/O Benchmark to Parallel NetCDF
- Results on IBM SP at SDSC
- Preliminary numbers - further optimization planned
- Parallel NetCDF provides a useful subset of HDF5 features that are more amenable to parallel I/O

PVFS and ROMIO

- Parallel Virtual File System
  - Parallel File System for Linux clusters
  - SDM work extending PVFS capabilities to better match scientific application requirements (structured file access)
- ROMIO
  - MPI-IO implementation used on most platforms
  - SDM work
    - Harnessing new capabilities in PVFS
    - Enabling access to grid I/O resources via MPI-IO interface
    - Extending hints available for performance tuning

Automated Hint Generation

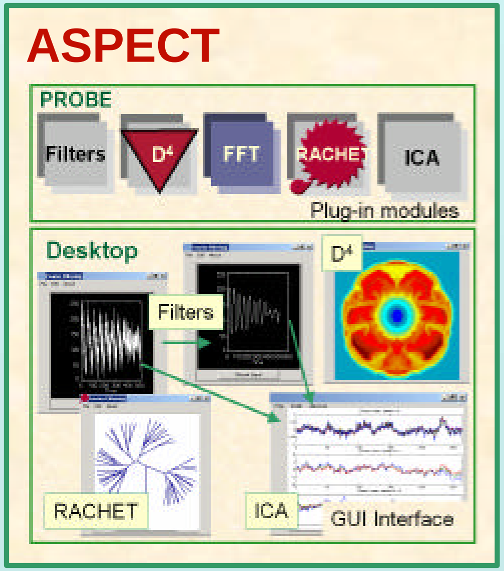
- Hints to underlying system software can boost performance by creating new pre-fetching and caching strategies
- “Right” set of hints varies by user, application, and system
- Automated generation of hints from I/O log files can lead to discovery of “most desirable” set of hints
- Feedback from previous hints to I/O system can shorten the discovery process
- Utilizing and extending MPI-IO hint mechanism is an ideal candidate for
- this approach



ASPECT: Adaptable Simulation Product Exploration & Control Tool

Application Scientists:

- ❑ Tony Mezzacappa, ORNL, Astrophysics
- ❑ David Erickson, ORNL, Climate
- ❑ John Drake, ORNL, Climate



Development Team:

- ❑ **Nagiza F. Samatova Project Lead**
- ❑ **George Ostrouchov**
- ❑ **Ian Watkins**
- ❑ **David Bauer**
- ❑ **Guru Kora**
- ❑ **Hoony Park**
- ❑ **Jennifer Golek**
- ❑ **Faisal AbuKhzam**
- ❑ **Yongming Qu**

Technology Collaborators:

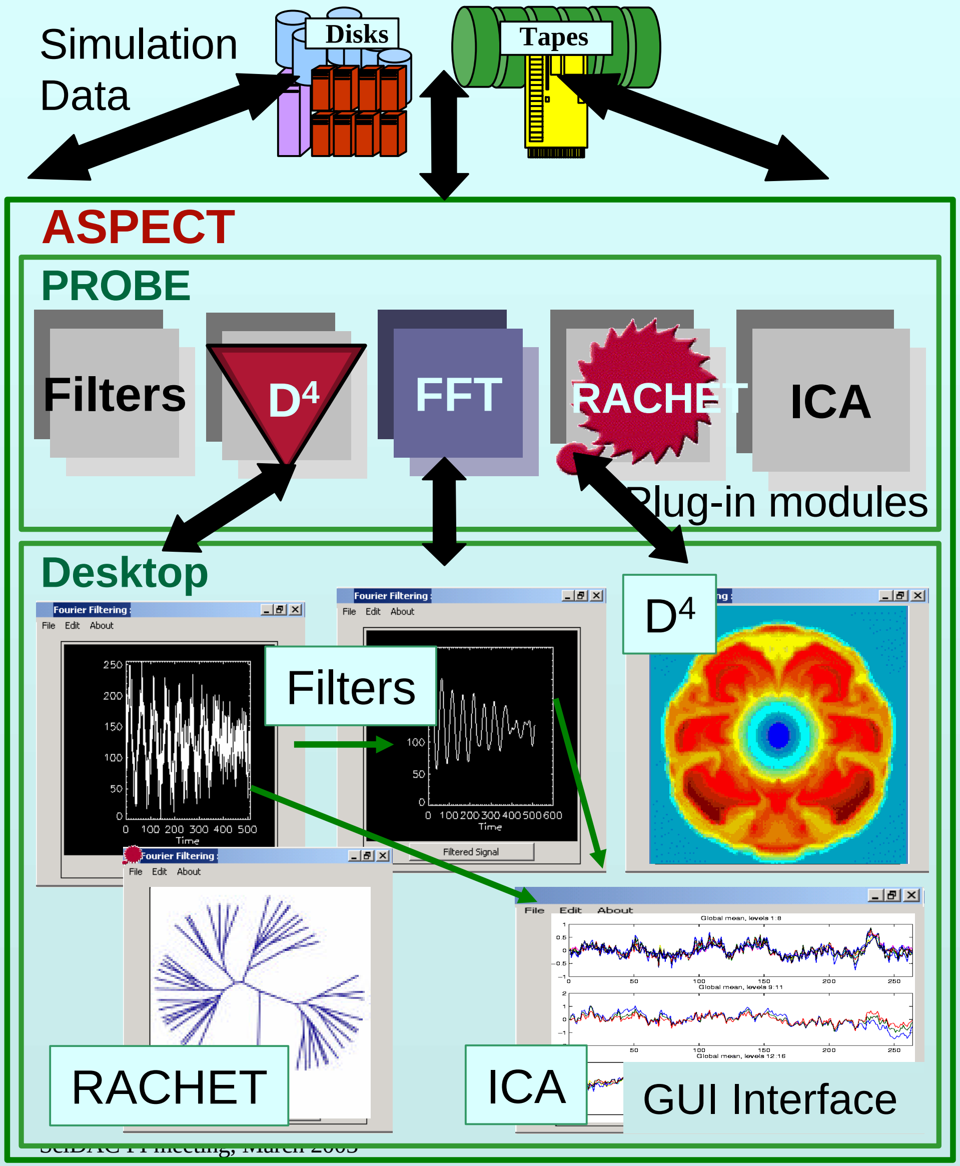
- ❑ Randy Burris, ORNL
- ❑ Ross Toedte, ORNL
- ❑ Rob Ross, ANL
- ❑ Bill Gropp, ANL
- ❑ Rajeev Thakur, ANL
- ❑ Rob Grossman, UIC
- ❑ Alok Choudhary, NWU
- ❑ Wei-keng Liao, NWU
- ❑ Jim Ahrens, LANL
- ❑ Gene Golub, Stanford U.
- ❑ Mike Langston, UTK

http://www.scidac.org/SDM/ASPECT

Typical Simulation Exploration Scenarios Driven by limitations of existing technologies

- **Post-processing Scenario:**
  - ⇓ Submit a long-running simulation job (weeks – months)
  - ⇓ Periodically check the status (run "tail -f" command on each machine)
  - ⇓ Analyze large simulation data set
- **Real-time Scenario:**
  - ⇓ Instrument a simulation code to visualize a field(s)
  - ⇓ While running a simulation job
    - Monitor the selected field(s)
    - If can not monitor, then either Stop a job or Continue running without monitoring and ability to view later what has been skipped
  - ⇓ If changing a set of fields to monitor, then go to 1

Improvements through — ASPECT
Data stream - not simulation - monitoring tool



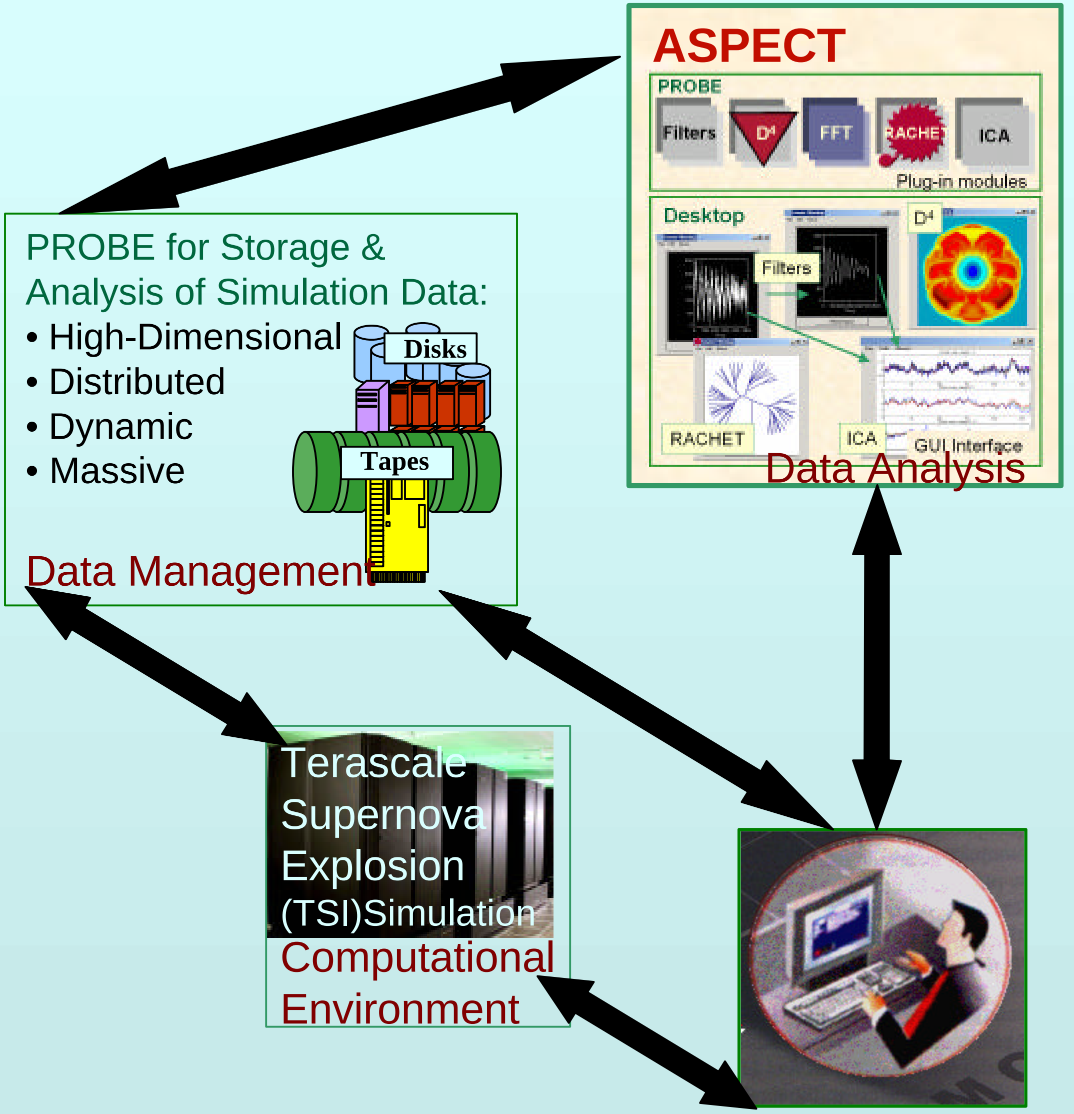
ASPECT's advantages:

- No simulation code instrumentation
- Single data — multiple views of data
- No interference w/ simulation
- Decoupled from the simulation

ASPECT's drawbacks:

- No computational steering
- No collaborative visualization

“Run and Render” Simulation Cycle



ASPECT

Data Analysis

PROBE for Storage & Analysis of Simulation Data:

- High-Dimensional
- Distributed
- Dynamic
- Massive

Data Management

Terascale Supernova Explosion (TSI)Simulation

Computational Environment

Application Scientist

ASPECT Features:

✓ Enables effective and efficient monitoring of data generated by long running simulations.

✓ Provides the GUI interface to a rich set of pluggable data analyses.

✓Supports scientific & statistical data analysis visualization via pVTK.

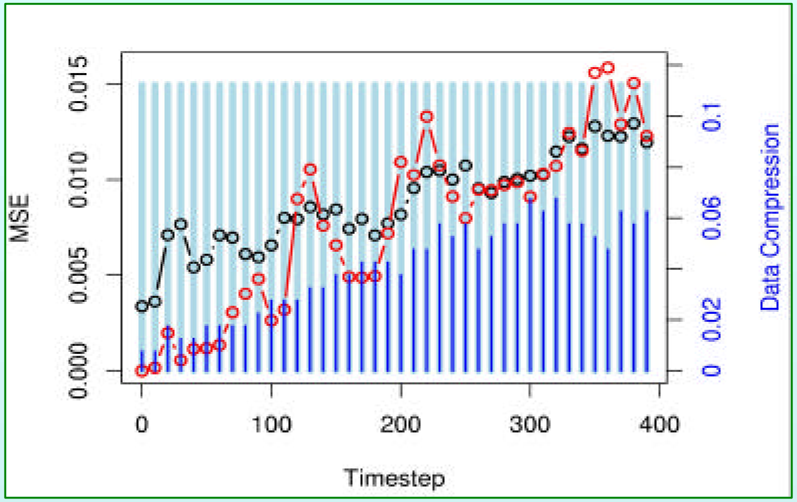
✓Handles large data sets via data reduction & parallel algorithms

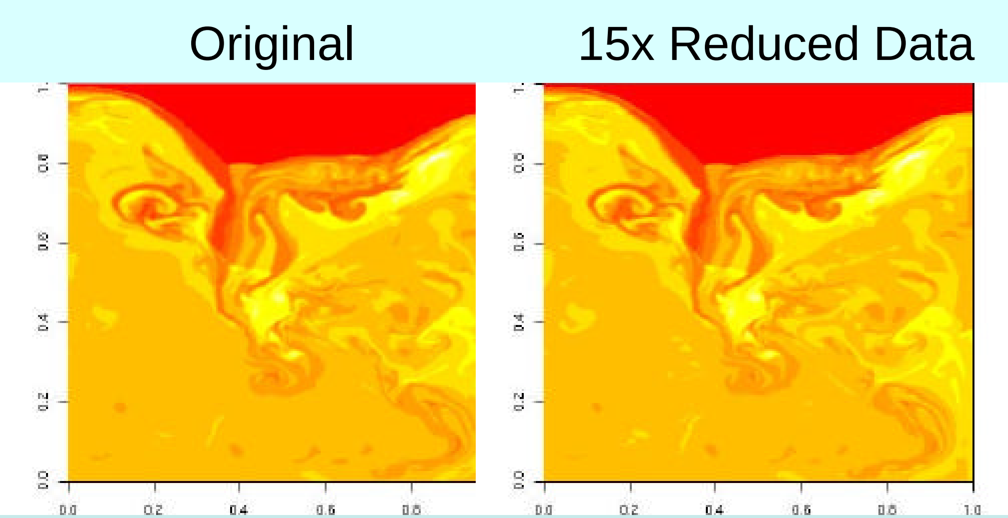
✓Provides efficient I/O through MPI-IO to NetCDF and HDF.

✓ Transfers data efficiently through UDB-based Sabul protocol.

Adaptive Data Reduction in ASPECT

- **SciDAC TSI simulation:**
  - **15 to 200 times reduction per time step**
  - **Outperforms sub-sampling 3 times for comparable MSE over all time steps**
  - **Provides 30-fold compression with 99% accuracy (captured variability)**
- **Based on SVD of contiguous field blocks**
- **Exploits spatial correlation & adapts to complexity of spatial field**
- **Parameter controls captured percentage of variation**
- **Linear time field restoration**




Original

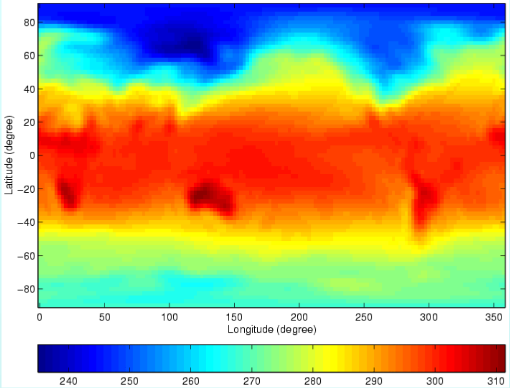
15x Reduced Data

Time step 390

This work is expanding the scientific understanding of global climate change

- **Goal**: understand changes in global temperatures
- **Problem**: separation of different sources, such as El Niño (ENSO) and volcano eruptions
- **Results**: first to identify ENSO component in zonally averaged re-analysis climate data from NCEP
- **Next steps**: refine techniques and scale up algorithms to identify more sources in other datasets
- **Collaborations**: vital to our success
  - Dr. B. D. Santer: climate expert
  - SciDAC team: latest computational methods
- **Other benefits**: general purpose software tools for dimension reduction and source separation; re-usable in other domains; synergistic to efforts at ORNL and LBNL
- **Work performed by**: Imola Fodor, Chandrika Kamath (LLNL)

New Algorithm Successfully Isolates the ENSO Component in Global Temperatures

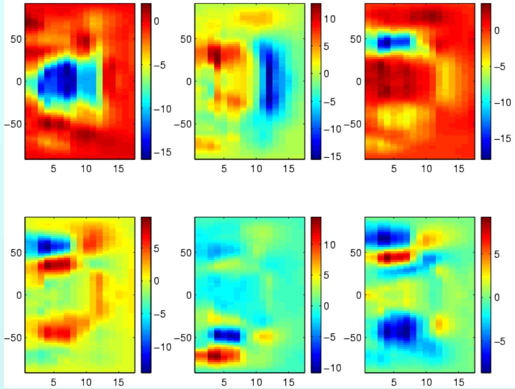


Raw data: 264x144x73x17

time lat lon altitude

Dimension reduction

PCA

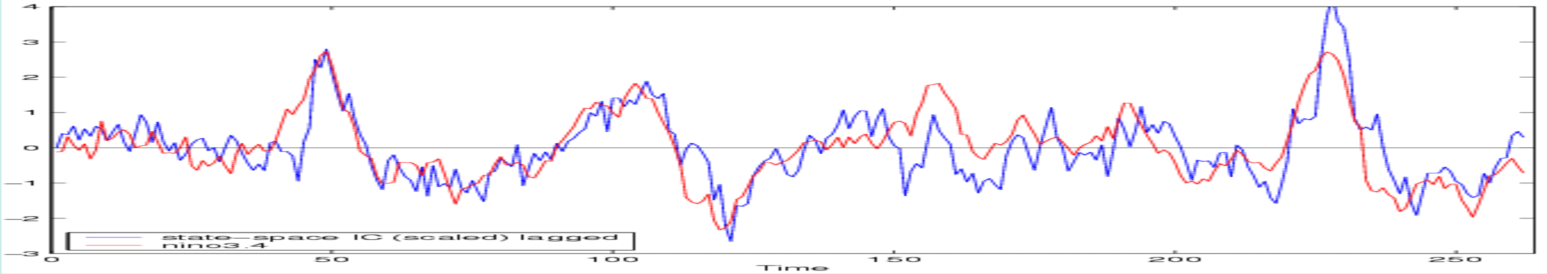


PC basis: 22x73x17

# reduced dimension

Source separation

ICA



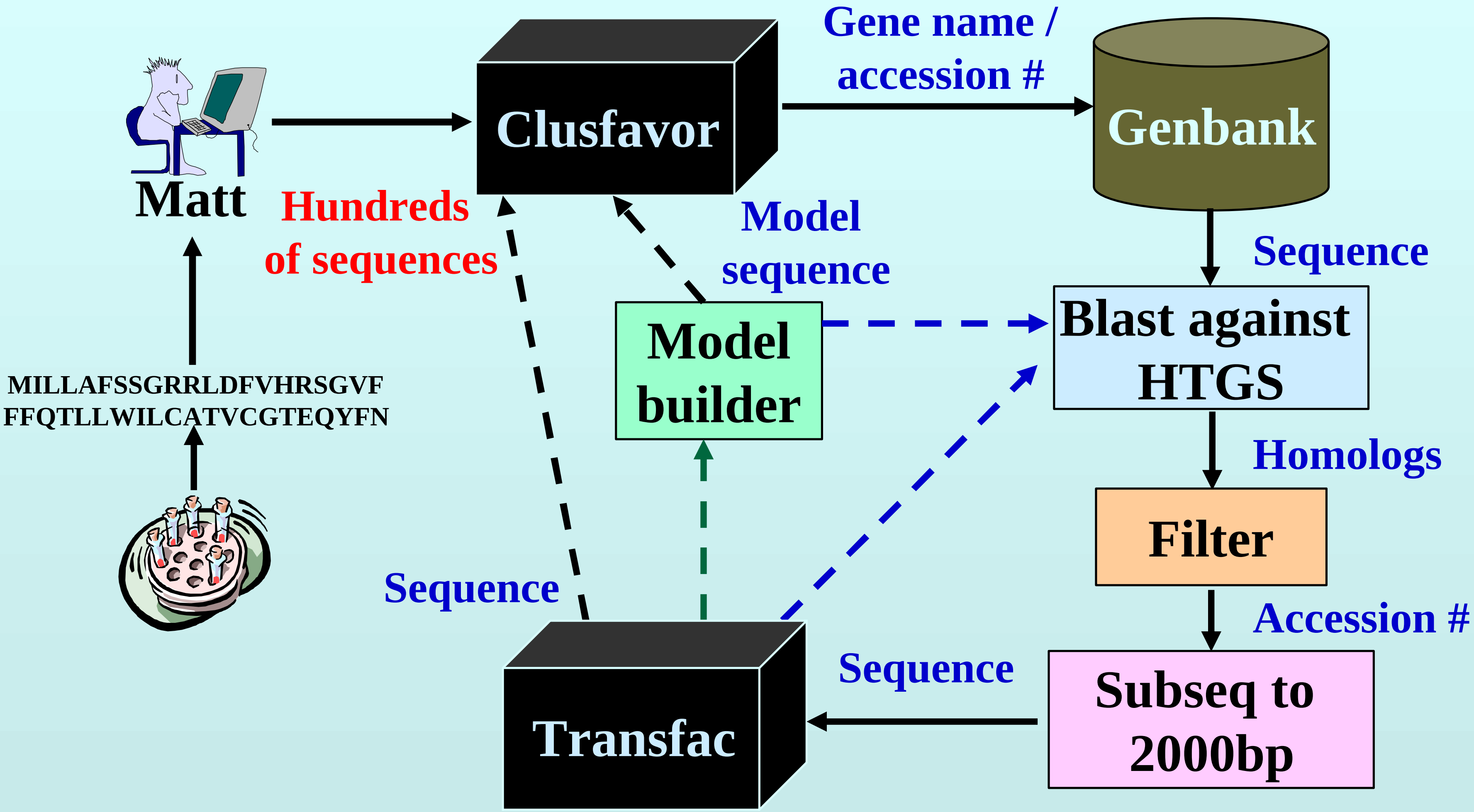
Estimated ENSO source component and Nino 3.4 ENSO index: 264x1

The excellent match indicates the success of our approach.

Distributed, Heterogeneous Data Integration

- Team members
  - Terence Critchlow (LLNL)
  - Calton Pu, Ling Liu (Georgia Tech)
  - Bertram Ludaescher, Amarnath Gupta, Ilkay Altintas (SDSC)
  - Mladen Vouk, Donald Bitzer, Munindar Singh, David Rosnick (NCSU)
  - Matt Coleman (LLNL)
- Helping scientists perform the complex data manipulations they need to perform their research

Motivating Use Case: Identifying Model Sequences



Matt

Clusfavor

Gene name / accession #

Genbank

Hundreds of sequences

Model sequence

Sequence

Model builder

Blast against HTGS

MILLAFSSGRRLDFVHRSGVF
FFQTLLWILCATVCGTEQYFN

Homologs

Filter

Accession #

Sequence

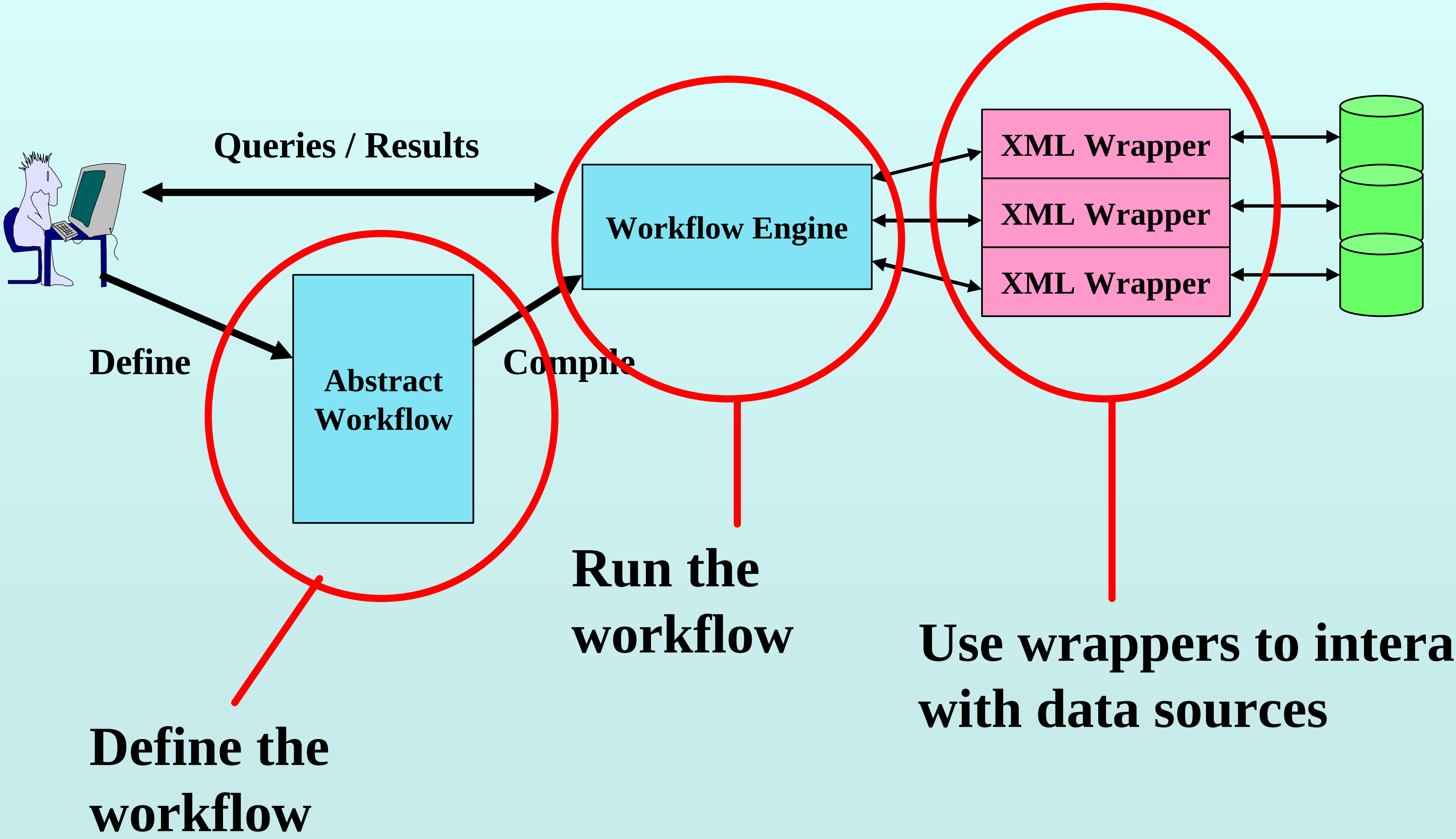
Sequence

Transfac

Subseq to 2000bp

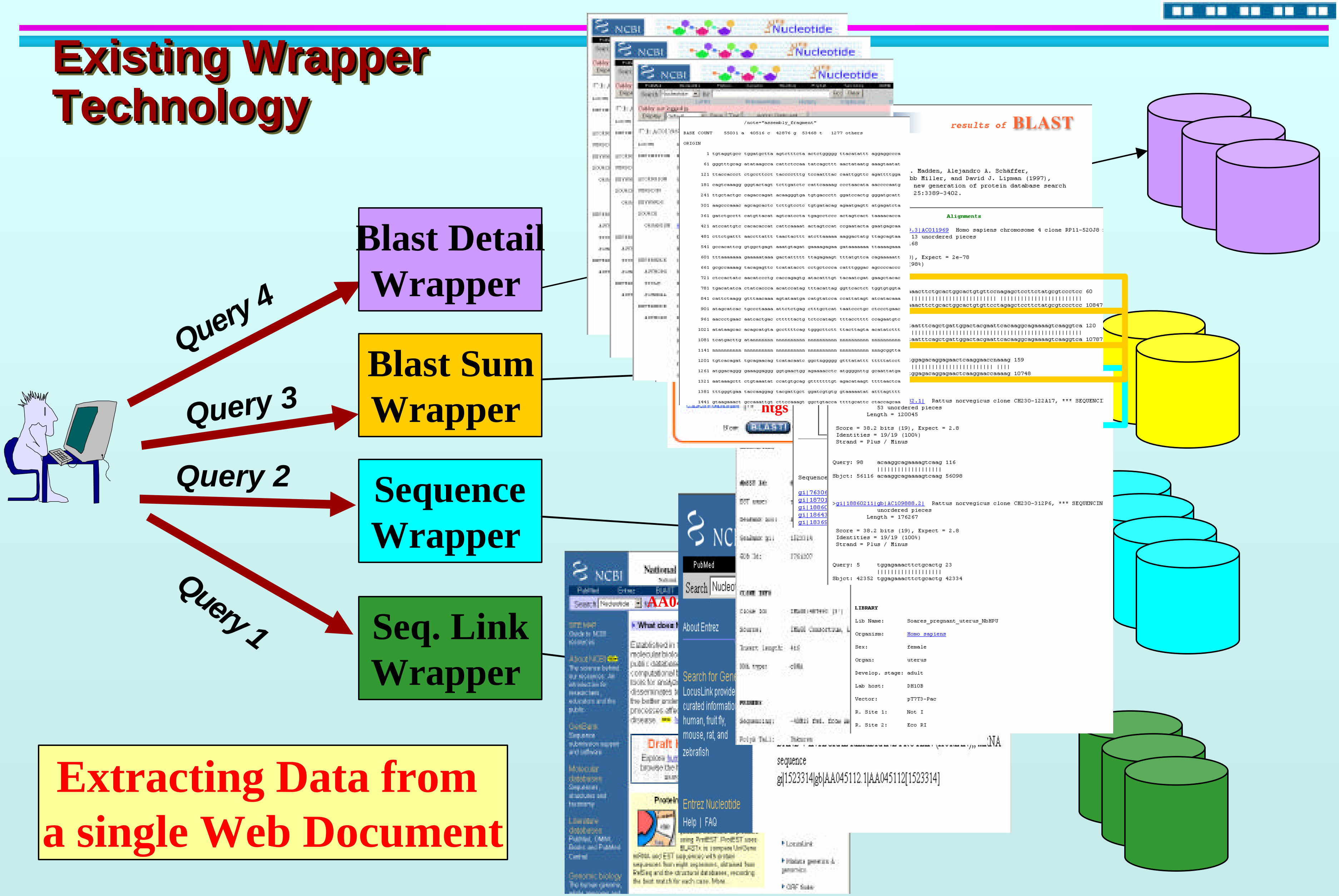
User View

Workflow based approach to data integration



SDM Enabling Technology: XWRAPComposer

Existing Wrapper Technology



Blast Detail Wrapper

Blast Sum Wrapper

Sequence Wrapper

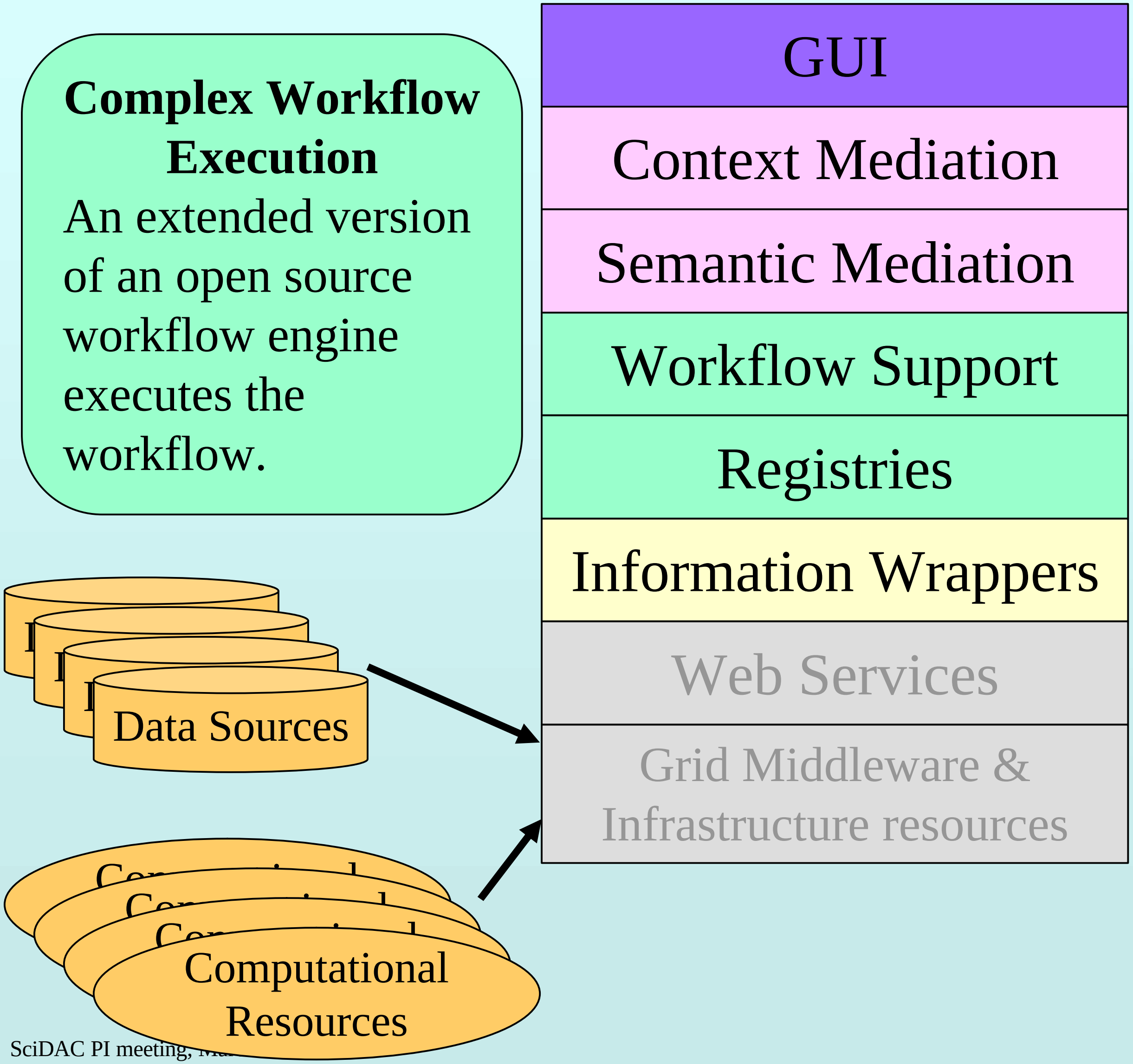
Seq. Link Wrapper

Extracting Data from a single Web Document

Architecture

Complex Workflow Execution

An extended version of an open source workflow engine executes the workflow.



Abstract Workflow Definition

Domain specific transformations use semantic mediation to map the abstract workflow into an executable format.

Automatic Wrapper Generation

The workflow engine is isolated from the data sources by wrappers.

Future Work

- Semantic mediation
  - Define abstract workflow language
  - Compile abstract workflow to executable workflow
    - Targeting XPDL (workflow engine input format)
- Extend open source workflow engine
  - Integrate open source workflow engine with existing workflow design tool
- Wrapper generation
  - Automatically define WSDL and SOAP interfaces

A prototype is being used by Matt Coleman at LLNL

FastBit: Compressed BitMap Index

Applied to:
HENP
Combustion

John Wu, Arie Shoshani, Doron Rotem

LBNL

HENP Collaborators: Jerome Lauret, STAR-BNL

Wei-Ming Zhang, Kent State University

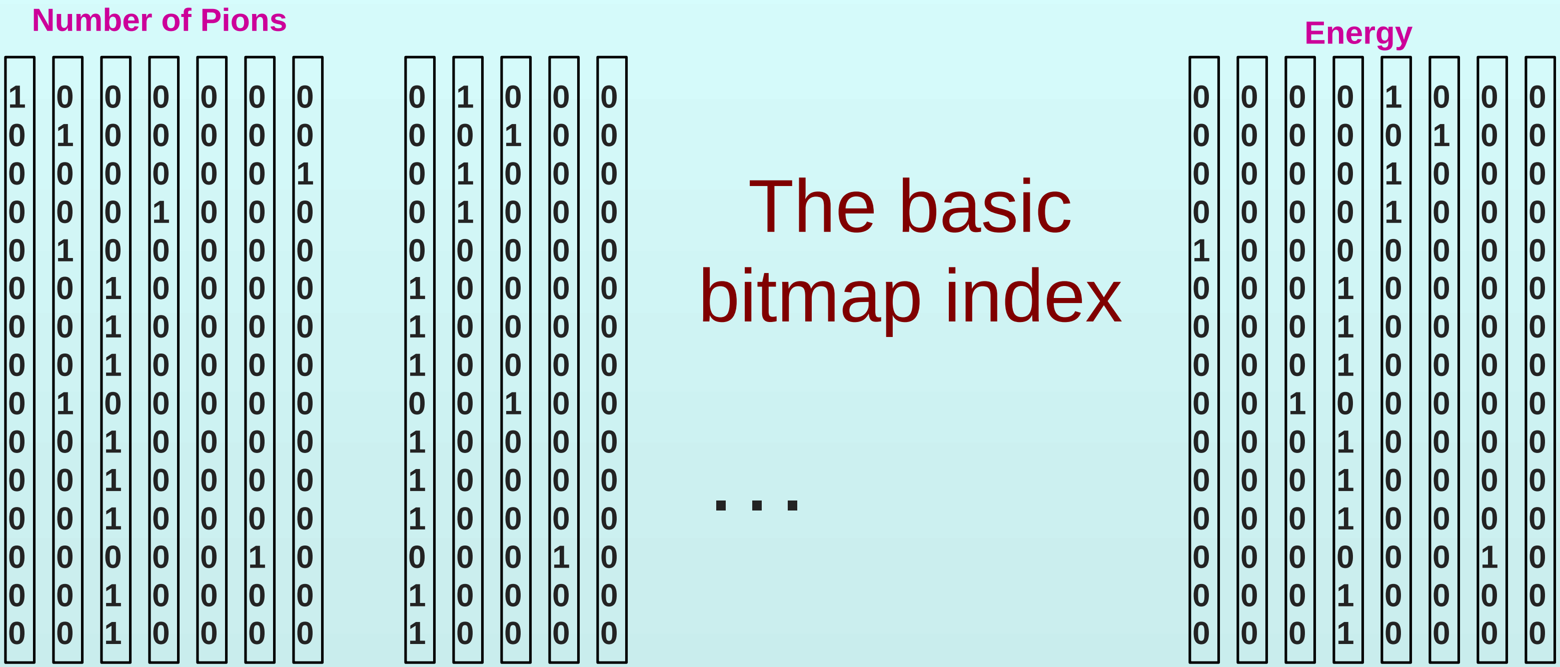
Combustion Collaborators: Wendy Koegler,
Jackie Chen, SNL

FastBit Index

- Need to search over
  - millions of objects (100s million events)
  - Hundreds of searchable attributes
- Most users specify range conditions on a handful of attributes
- Our method (FastBit) is effective with high cardinality attributes by applying:
  - Binning the attribute values
  - effective compression of bitmaps
  - Optimize Compression for Computation Efficiency

Basic Bitmap Index

- Bitmap index idea: have one bitmap for each value
- For scientific data: values are numeric and have large cardinality
- partition into bins (e.g. 100): 0-10, 11-20, …



Number of Pions

1	0	0	0	0	0	0
0	1	0	0	0	0	0
0	0	0	0	0	0	1
0	0	0	1	0	0	0
0	1	0	0	0	0	0
0	0	1	0	0	0	0
0	0	1	0	0	0	0
0	0	1	0	0	0	0
0	1	0	0	0	0	0
0	0	1	0	0	0	0
0	0	1	0	0	0	0
0	0	1	0	0	0	0
0	0	0	0	0	1	0
0	0	1	0	0	0	0
0	0	1	0	0	0	0

0	1	0	0	0
0	0	1	0	0
0	1	0	0	0
0	1	0	0	0
0	0	0	0	0
1	0	0	0	0
1	0	0	0	0
1	0	0	0	0
0	0	1	0	0
1	0	0	0	0
1	0	0	0	0
1	0	0	0	0
0	0	0	1	0
1	0	0	0	0
1	0	0	0	0

The basic bitmap index

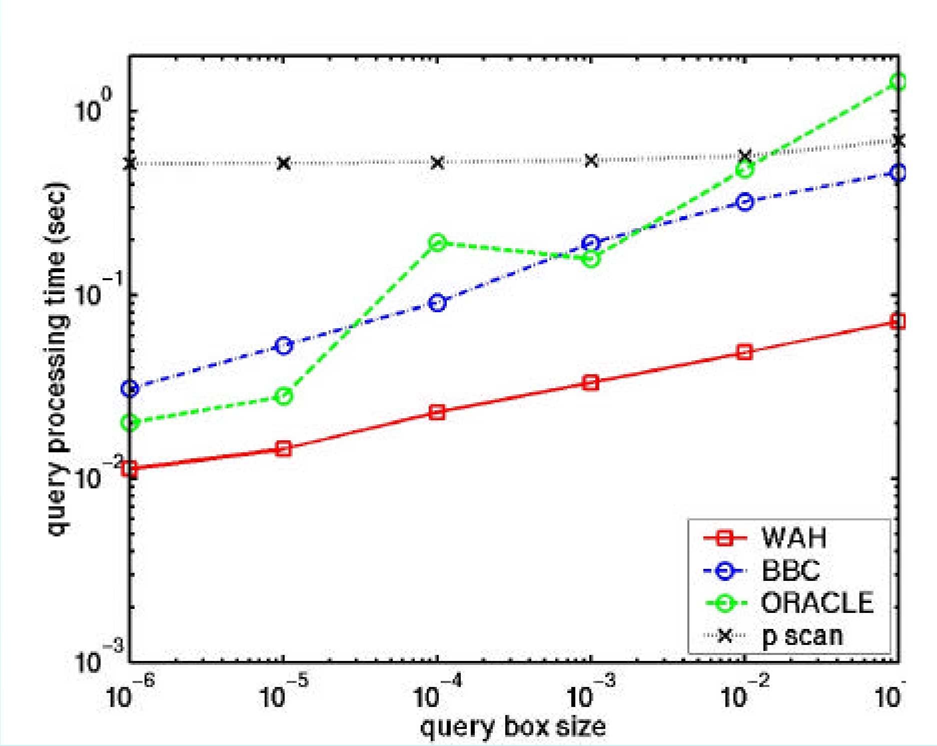
. . .

Energy

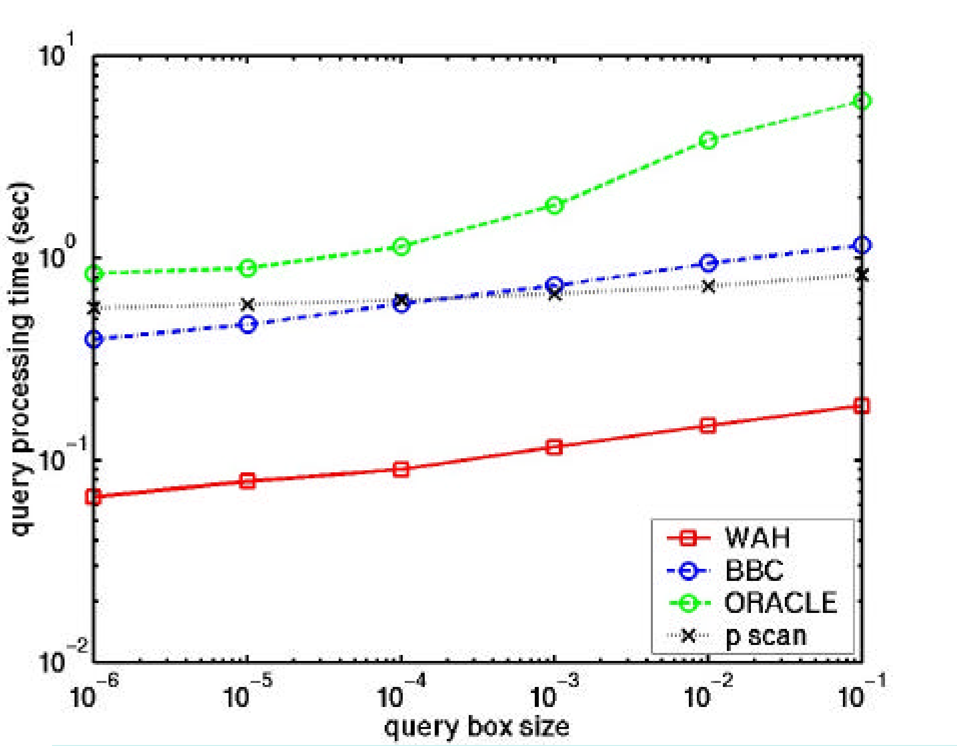
0	0	0	0	1	0	0	0
0	0	0	0	0	1	0	0
0	0	0	0	1	0	0	0
0	0	0	0	1	0	0	0
1	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0
0	0	0	1	0	0	0	0
0	0	0	1	0	0	0	0
0	0	1	0	0	0	0	0
0	0	0	1	0	0	0	0
0	0	0	1	0	0	0	0
0	0	0	1	0	0	0	0
0	0	0	0	0	0	1	0
0	0	0	1	0	0	0	0
0	0	0	1	0	0	0	0

- Only a single "1" in each row
- WAH: Optimize Compression for Computation Efficiency

Performance of Partial Range Queries
2.2 million records of STAR data



2 attributes per query



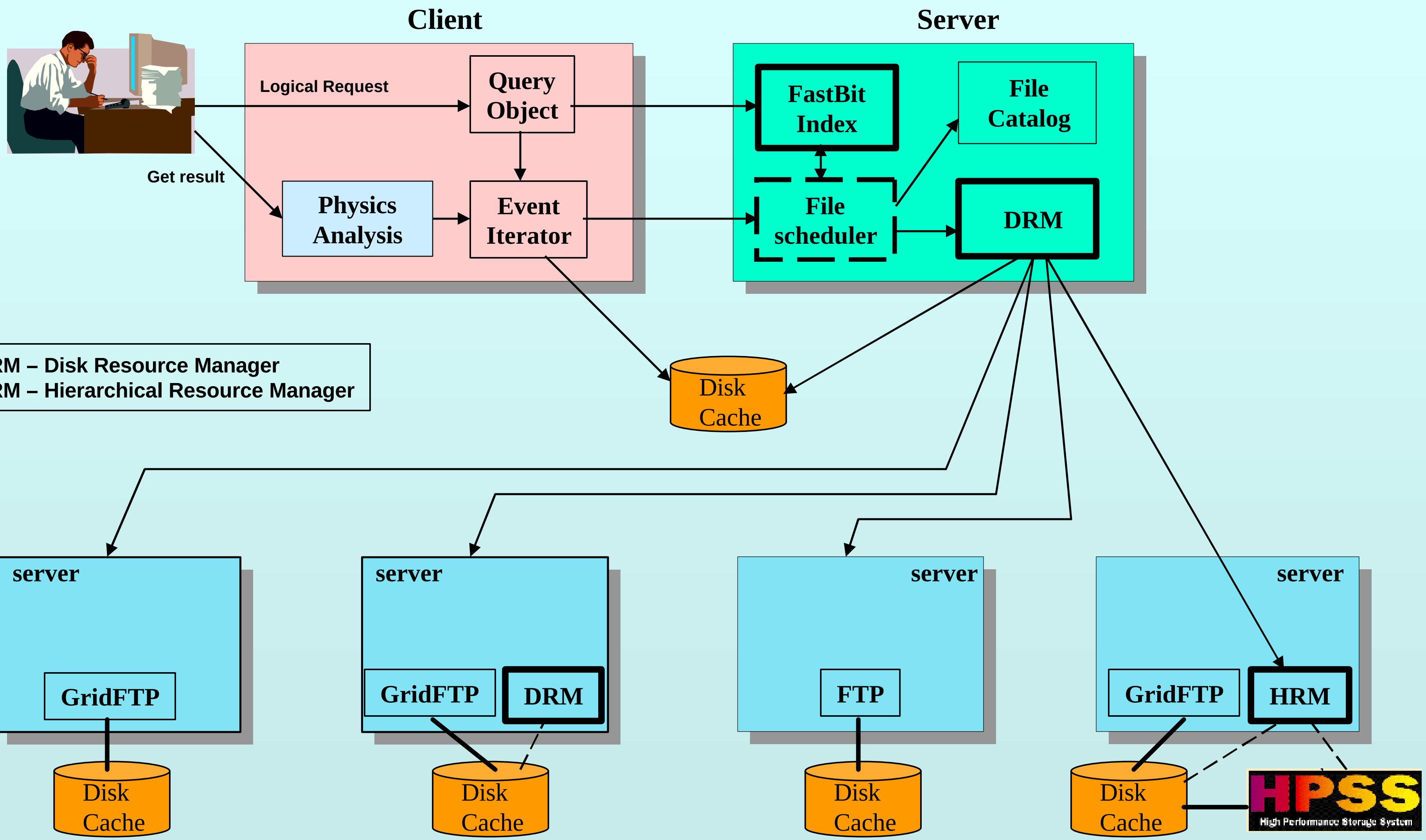
5 attributes per query

- **WAH compressed indexes are**
  - **>10X faster than ORACLE,**
  - **>5X faster than our implementation of BBC (BBC is also used in ORACLE)**
- **12 most queried attributes are used, average attribute cardinality 222,000**

Using Bitmap Indexing in the STAR Analysis Framework

- **Generate large amounts of raw data**
  - Collect from experiments or large simulations
  - STAR: ~ 100 Million collision events a year
  - STAR: Each event ~1-5 MB
- **Post-processing of data**
  - process data (find particles produced, tracks)
  - generate summary data
    - e.g. momentum, no. of pions, transverse energy
    - Number of properties is large (50-100 attributes)
- **Analyze data**
  - use summary attributes to select relevant events
  - extract subsets from the large dataset
    - Need to access events based on partial properties specification (range queries)
    - e.g. ((0.1 < Energy < 0.2) ^ (10 < Np < 20)) v (N > 6000)
  - Current practice: generate pre-selected subsets (called micro-DST)

Architecture for Dynamic Analysis Framework



DRM – Disk Resource Manager
HRM – Hierarchical Resource Manager

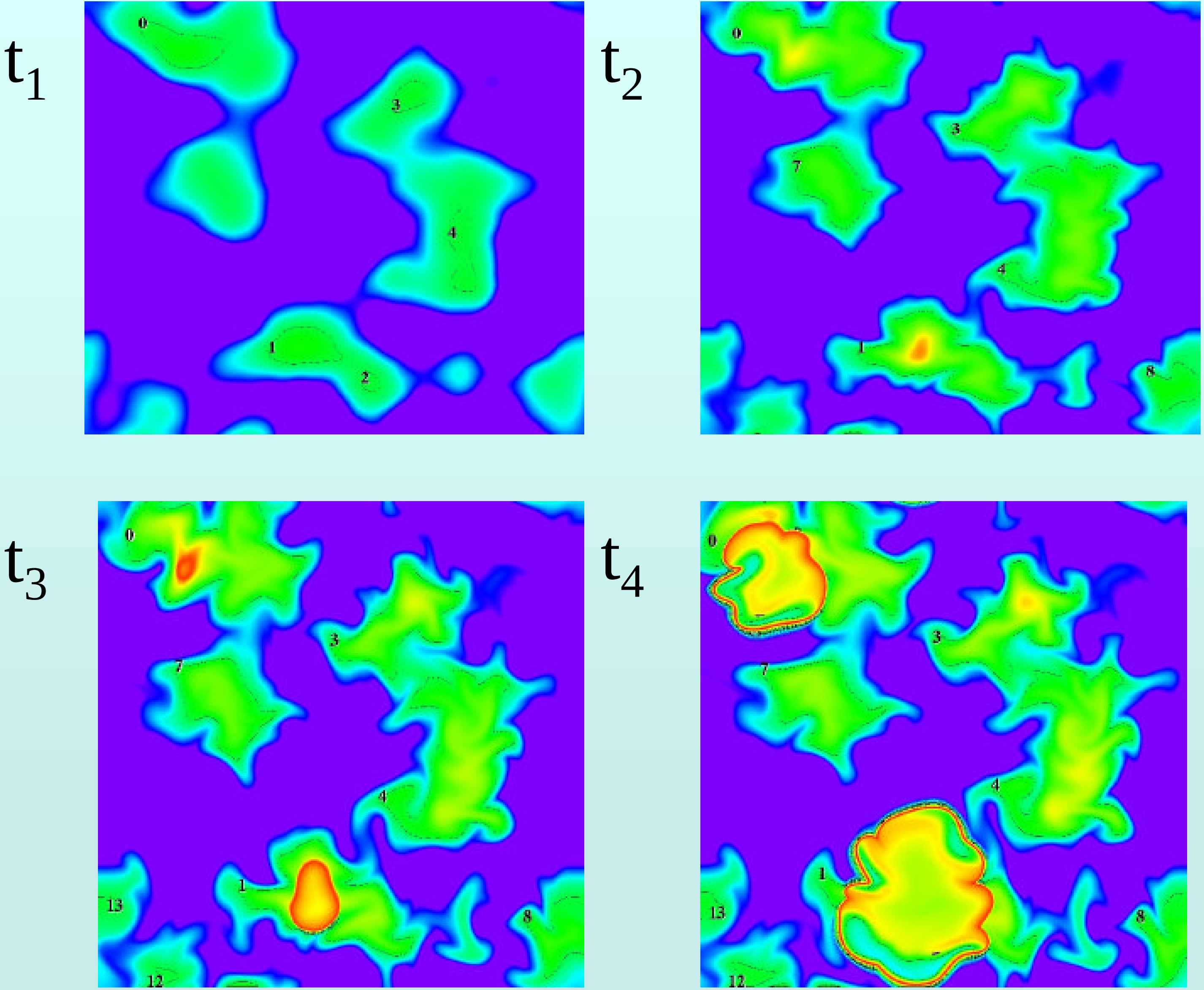
Discovering Combustion Flame Fronts using Bitmap Indexing Technology

- **Characteristics of feature tracking in combustion data analyses**
  - High-fidelity simulation on 2D or 3D uniform grids, say, 1000 x 1000 x 1000, for hundreds of time steps
  - A model of hydrogen-air mixture has a dozen attributes per grid point, a realistic model has many more
  - Features are defined to be regions that satisfy the user specified conditions, such as,
    "600 < Temperature < 700 AND HO_2 concentration > 10^{-7}"
- **Goal:**
  - Interactive feature tracking
  - Multiple attribute conditions for specifying regions
- **We use FastBit for:**
  - Cell identification
  - Region growing
  - Region tracking

Tracking Features Over Time Steps

- Features are outlined with thin black lines

- Red color indicates a high concentrate of a transient chemical HO_2



Performance- Feature Tracking

- Feature tracking in interactive time -- the time is less than ten seconds in most cases
- Feature identification (searching and region growing) in one time step on 1000^3 grid may be completed in less than ten seconds
- Times are for all time steps

Average time (seconds) to perform steps of feature tracking

600 x 600 grid, 69 time steps (795 MBs)				**1344 x 1344 grid, 335 time steps (19.3 GBs)**			
# attr	Search	Grow	Track	# attr	Search	Grow	Track
1	1.06	0.22	0.02	1	5.71	2.05	0.12
2	1.67	0.17	0.01	2	7.39	1.24	0.12
3	2.12	0.14	0.01	3	8.92	0.58	0.11
4	2.62	0.14	0.01	4	10.30	0.47	0.10

Adaptive File Caching and Replication in Distributed Systems

Goals

- Develop a coordinated optimal file caching and replication of distributed datasets

Two Principal Components of Policy Advisory Module

- A disk cache replacement policy
    - Evaluates which files are to be replaced when space is needed
- Admission policy for file requests
    - Determines which request is to be processed next
    - e.g. may prefer to admit requests for files already in cache

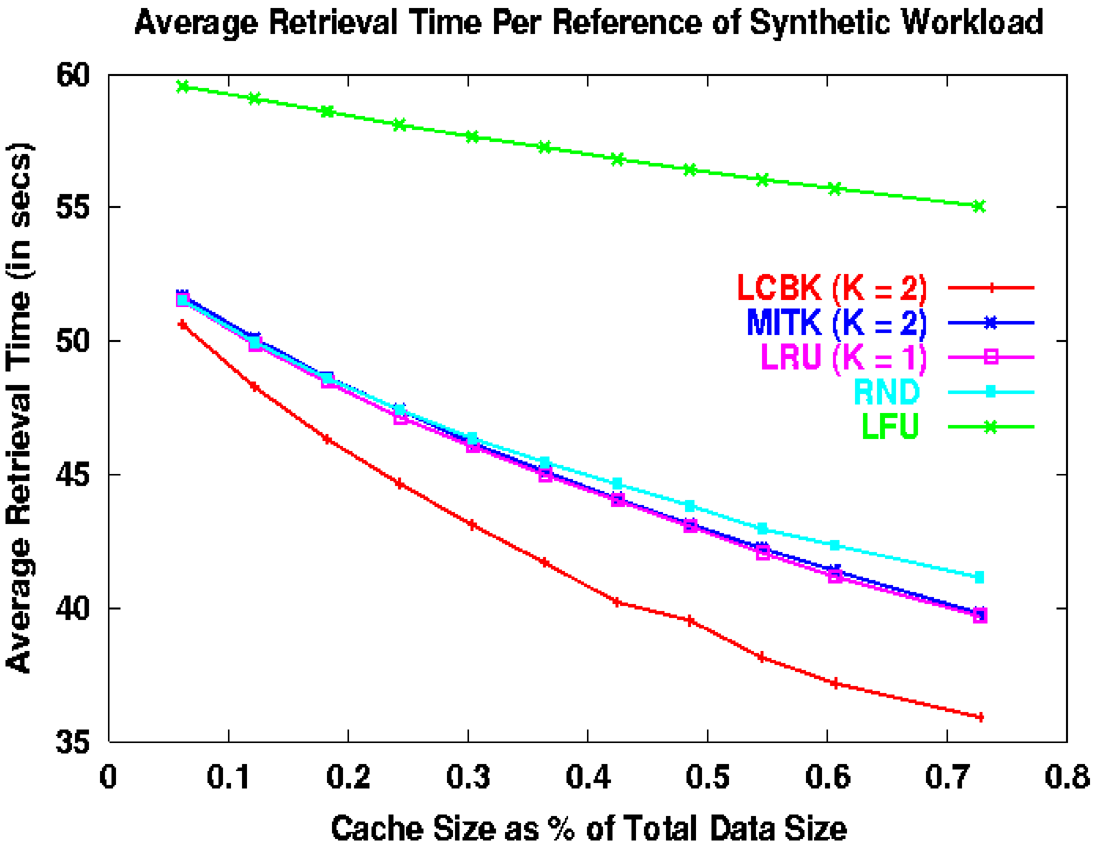
Work Perfomed by:

- Ekow Otoo, Doron Rotem, Arie Shoshani (LBNL)

Some Simulation Results

- LCB-K is the winner
- MIT-K, LRU and RND comparable
- LFU pretty bad

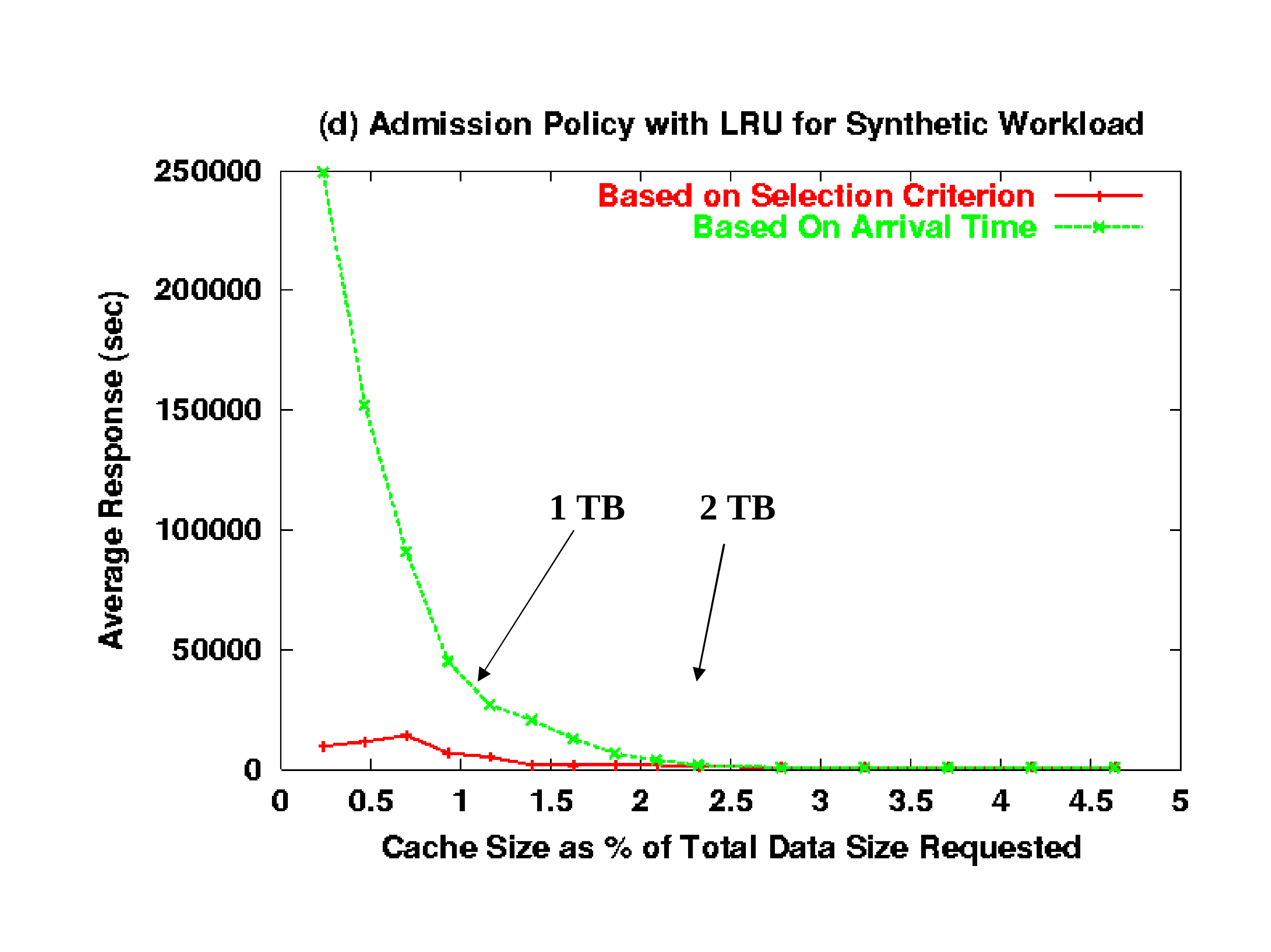
•Lower values represent better policies



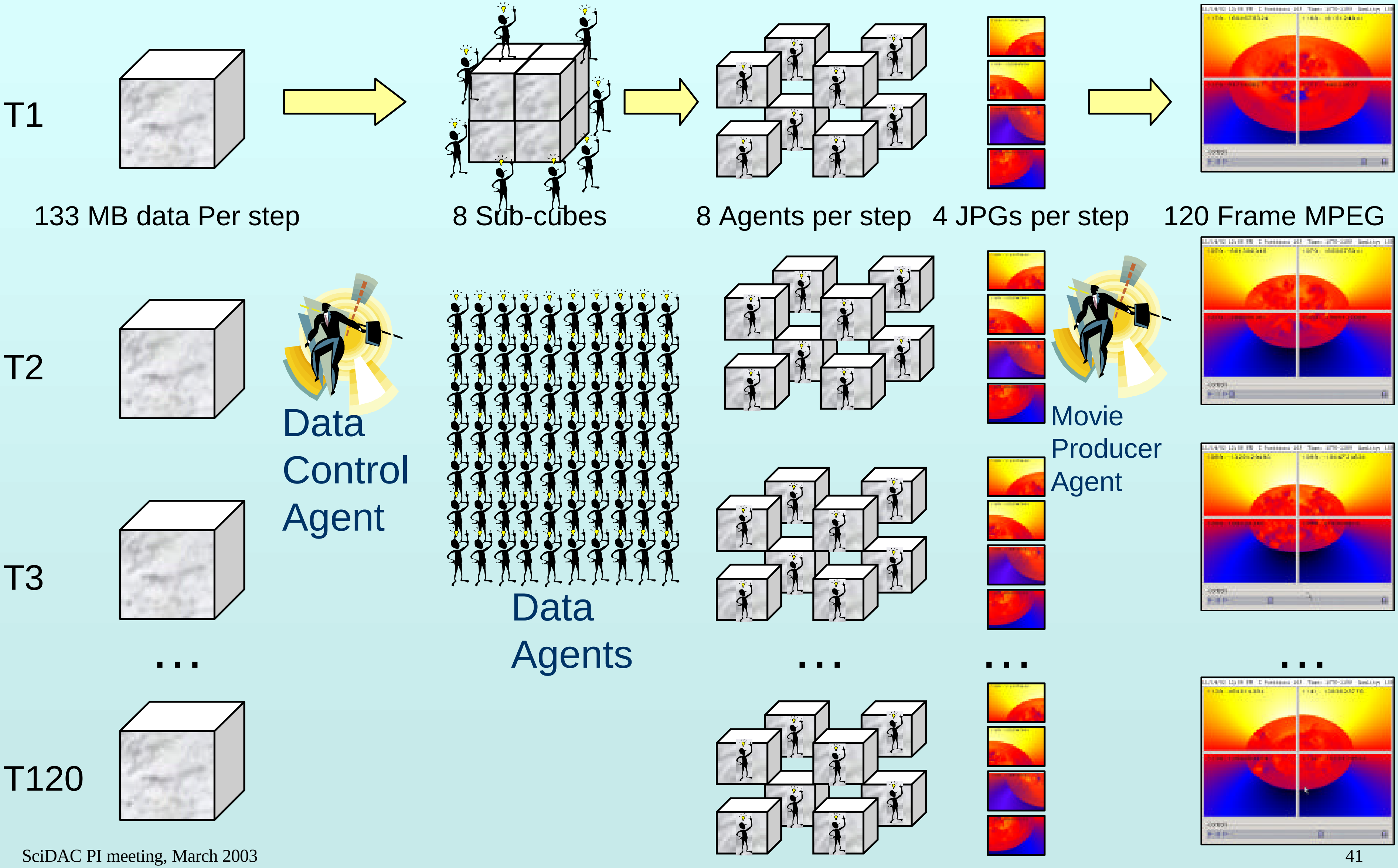
<u>Replacement Policies:</u>

•RND: Random
•LFU: Least Frequently Used
•LRU: Least Recently Used
•MIT-K: Maximum Inter-Arrival Time based on last K references
•LCB-K: Least Cost Beneficial based on last K references

Admission Policy Simulation



A Multi-agent System for Analyzing Massive Scientific Data



T1

133 MB data Per step

8 Sub-cubes

8 Agents per step

4 JPGs per step

120 Frame MPEG

T2

T3

. . .

T120

Data Control Agent

Data Agents

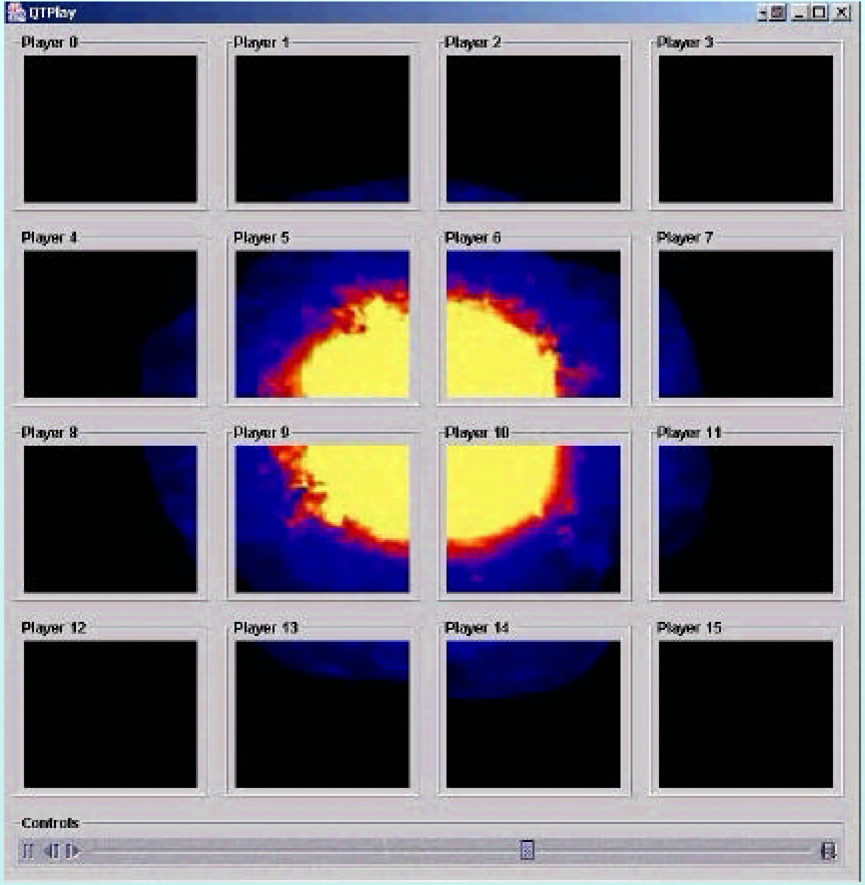
Movie Producer Agent

. . .

. . .

. . .

Results:

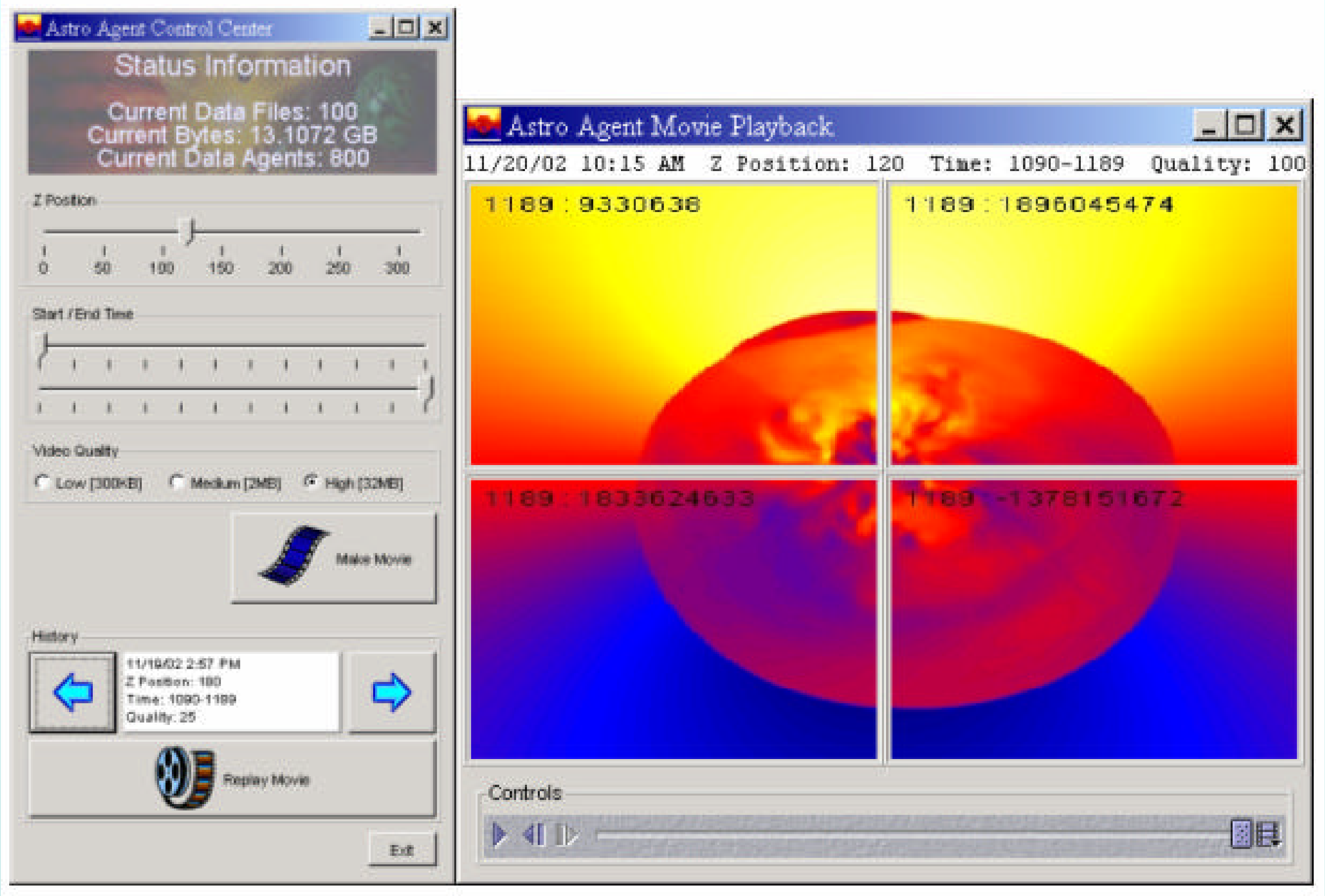


- Astrophysics data from 194 files
- 800 agents, each deployed to create an operational picture for 100 time steps
- Running on three machines
- Agents automatically created every time a new simulation output file is created

Work performed by:

Thomas Potok, Joel Reed, Tony Mezzacappa, John Blondin, Rick Sheldon

Controlling the movie



Astro Agent Control Center

Status Information

Current Data Files: 100
Current Bytes: 13.1072 GB
Current Data Agents: 800

Video Quality

○ Low [300KB] ○ Medium [2MB] ● High [32MB]

History

11/19/02 2:57 PM
Z Position: 180
Time: 1090-1189
Quality: 25

Astro Agent Movie Playback

```
11/20/02 10:15 AM  Z Position: 120  Time: 1090-1189  Quality: 100
```

Work with individual application scientists

- Close collaboration with individuals
  - Matt Coleman - LLNL (Biology)
  - Tony Mezzacappa – ORNL (Astrophysics)
  - Ben Santer – LLNL
  - John Drake - ORNL (Climate)
  - Doug Olson - LBNL, Wei-Ming Zhang – Kent (HENP)
  - Wendy Koegler, Jacqueline Chen – Sandia L. (Combustion)
  - Mike Papka - ANL (Astrophysics Vis)
  - Mike Zingale – U of Chicago (Astrophysics)
  - John Michalakes – NCAR (Climate)

Re-apply techniques to new applications

- Parallel NetCDF
  - Astrophysics → Climate
- Adaptive data reduction
  - Astrophysics → Climate
- Compressed bitmaps
  - HENP → Combustion
- Robust File Replication
  - HENP → Climate
- Signal Separation
  - Climate → Fusion

Summary

- Our guiding principles served us well
- We are focused, result oriented
- Technology migration to new applications
- Clear future path, lots to do
- Our focus: getting technology into the hands of scientists